GREASE II

Medidas de Semelhança Geográfica

Daniel Amoedo
damoedo@lasige.fc.ul.pt

Objectivo

- Semelhança entre sumários geográficos (Geo Ontology).

- Qual o melhor algoritmo de semelhança semântica?

Daniel Amoedo
damoedo@lasige.fc.ul.pt

Semelhança Semântica

- Medidas capazes de calcular semelhança entre 2 termos de uma ontologia.
- Information Content (based on the frenquency of the term)
- GEO-NET-PT (DAG)
  - Frt_id = PRT

Daniel Amoedo
damoedo@lasige.fc.ul.pt

Semelhança Semântica

- Algoritmos Implementados
  - ssmResnick

$$sim_{Res}(c_1, c_2) = IC(c_{MICA})$$

  - ssmJiangConrath

$$dist_{Lin}(c_1, c_2) = 1 - sim_{Lin}(c_1, c_2) = \frac{IC(c_1) + IC(c_2) - 2 \times IC(c_{MICA})}{IC(c_1) + IC(c_2)}$$

  - ssmLin

$$sim_{Lin}(c_1, c_2) = \frac{2 \times IC(c_{MICA})}{IC(c_1) + IC(c_2)}$$

Daniel Amoedo
damoedo@lasige.fc.ul.pt

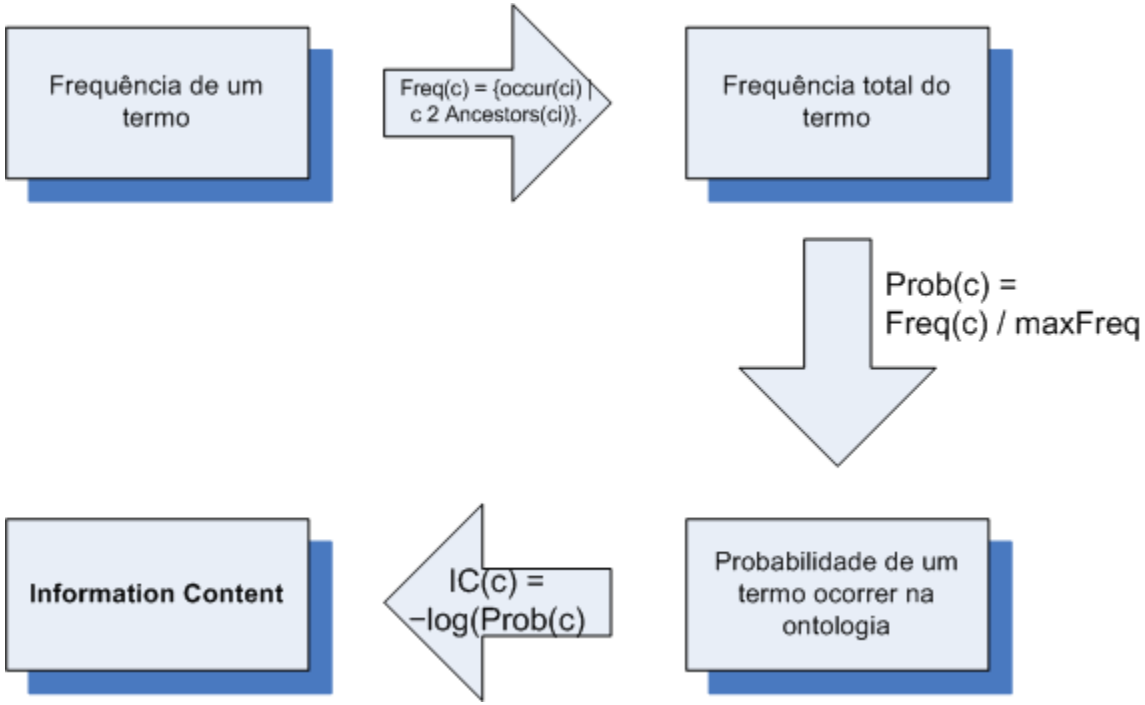
Semelhança Semântica

- MICA
  - Most Information Content Ancestor
- GRASM
  - Graph-Based Similarity Measure

Daniel Amoedo
damoedo@lasige.fc.ul.pt

Information Content

- Cálculo



Daniel Amoedo
damoedo@lasige.fc.ul.pt

Information Content

- Probabilidade de ocorrer um termo na ontologia

- Menor IC -> menos específico é um termo.
  - hfreq (topo) = hfreq(total) ⟹ Termo de menor IC

- Partilha de informação com o ancestor.
  - IC("Lisboa") = 3.067621
    - IC("augusta") = 5.5333347 (0.34%)
    - IC("janelas verdes") = 8.15579 (0.0008%)

Daniel Amoedo
damoedo@lasige.fc.ul.pt

Information Content

- Google N-Grams Corpus (Information Content)
  - Frequência observada de palavras , tal como nome de locais portugueses
  - 1 trilião de tokens de palavras tiradas de publicidade acessível através de páginas Web

	Geo-Net-Pt	Found in GNGC	Percentage
Unique Names	78070	65739	84.205%
Feature Names	199053	104982	52.741%

Daniel Amoedo
damoedo@lasige.fc.ul.pt

SSM Calculation

- ssm_graphpath
  - fid_1 , fid_2, distance
- Ssm_termfreq
  - Fid, freq, hfreq, prob, info_content, **rel_ic**

Daniel Amoedo
damoedo@lasige.fc.ul.pt

SSM

- Semelhança entre dois termos da ontologia
- Similarity Between Summarys.
  - Algoritmos:
    - MaxMin
    - MinMax
    - Média
    - Depth
- Best Match Summary
  - (search, n summarys, ssm, sbs) ➡ Best n summary

Daniel Amoedo
damoedo@lasige.fc.ul.pt

Geo Scope

- Saber o âmbito geográfico de um sumário geográfico.
- Heurísticas (ineficaz)
- Ancestor de menor IC, que contém pelo menos 1 termo filho de cada nome da pesquisa.

Daniel Amoedo
damoedo@lasige.fc.ul.pt

PROBLEMA ?

- Desambiguação de nomes
  - Search(“lisboa”) = 41 features.
  - Search(“porto”) = 60 features.
  - Procuramos um arruamento?
    - Termos cujo tipo não é “arruamento”.
    - Pesquisa específica para “arruamentos”.
      - Search(“Lisboa”) : “rua Lisboa” ? “Largo Lisboa” ?

Daniel Amoedo
damoedo@lasige.fc.ul.pt

Desambiguação de Nomes

- Encontrar o ambito geográfico do resumo
- Para cada nome, encontrar o termo na ontologia semânticamente mais parecido com o âmbito encontrado

Daniel Amoedo
damoedo@lasige.fc.ul.pt

Scope (Problemas)

- Problemas
  - Geo-Net-Pt
    - Calheta (conselho da Madeira)
      - ID 69 ⟹ n_name : “calheta (madeira)”
    - Calheta (Localização nos Açores)
      - ID 150296 ⟹ n_name : “calheta”
    - Nomes Alternativos
      - Nova Feature
      - Sem ligações aos filhos do ancestor
    - Zonas sem precedências
      - Ex: ‘id’ = 349114, n_name = ‘alfama’

Daniel Amoedo
damoedo@lasige.fc.ul.pt

Scope (Problemas)

- SSM
  - Visão hierarquica
    - Cega a nível espacial
  - Termos ao mesmo nível na ontologia
  - CommonAnc(84, 16) = CommonAnc(162,16) = [94, 418745]
  - SSM(84,16) < SSM (162,16)

F_id	T_id	N_name	hfreq
84	CON	Castro Verde	22804018
162	CON	Matosinhos	393344590
16	NT2	Algarve	5060343466

Daniel Amoedo
damoedo@lasige.fc.ul.pt

Testes

- Páginas www.portugaltribe.com
  - Páginas anotadas
- GeoScope

Pag Portugal Tribe	ID do scope	T_id scope	N_name scope
Porto e Norte de Portugal	196	NT2	norte
Lisboa e Vale do Tejo	418732	PRO	estremadura
Alentejo	12	NT2	alentejo
Algarve	17	NT3	algarve
Açores	252	NT3	regiao autonoma dos acores
Madeira	418745	PAI	portugal

Daniel Amoedo
damoedo@lasige.fc.ul.pt

Work in Progress

- Qual a melhor medida semântica!?
- Scope Penalty
  - Scope optimizado
    - Descartar Outliers
    - Referência na página a um local.
    - Nome não encontrado na geo-net-pt
  - Ganho
- Ferramenta Web
- Descriminar arruamentos (GeoScope)

Daniel Amoedo
damoedo@lasige.fc.ul.pt

Sumário

- Feito:
  - Implementação de algoritmos de semelhança semântica aplicado ao Geo Ontology.
  - Saber o âmbito de um resumo geográfico
  - Similaridade entre dois sumários geográficos
  - Saber dentro de um conjunto de sumários geográficos, o que mais se assemelha a uma pesquisa geográfica dada.

Daniel Amoedo
damoedo@lasige.fc.ul.pt

Sumário

- Em falta
  - Encontrar o melhor algoritmo aplicado à ontologia.
  - Alargar a pesquisa do âmbito geográfico aos arruamentos
  - Scope optimizado

Daniel Amoedo
damoedo@lasige.fc.ul.pt

FIM